

Research on Human Pose Estimation Based on 2D and 3D Classification

Xu Ji¹^a, Can Liu²^b and Lingyue Zeng^{3*}^c

¹Tianjin Institute of Software Engineering, Tianjin University of Technology, Tianjin, China

²School of Computing and Data Science, Xiamen University Malaysia, Kuala Lumpur, Malaysia

³Computer Science, University of Liverpool, Liverpool, U.K.

Keywords: Human Pose Estimation, Top-Down, Bottom-Up, Heatmap, Regression.


Abstract: Human pose estimation (HPE) refers to the use of computer vision technology and machine learning methods to detect and analyse human pose information from images or videos. The early mainly focused on solving single person pose estimation problems. It treats the problems as classification or regression problems using a global feature. However, such methods are not accurate enough and only suitable for simple scenarios. The application of HPE estimation based on deep learning is very promising and can provide many important advantages and innovations in related field. Therefore, this paper mainly focuses on several methods and technical advances for 2D and 3D HPE based on deep learning. This paper aims to introduce the technique and compare the characteristics of different categories. Besides, it covers various datasets since 2015 and different metrics for evaluating methods performance and illustrating considerable potential for future development. Finally, this paper discusses the advantages and limitations through comparison of methods. With the development of computer processing ability, the future human pose estimation tasks will make more progress in improving accuracy, expanding application scenarios and optimizing efficiency.


1 INTRODUCTION


Human pose estimation (HPE) involves refers to recognizing different parts of human body and constructing representations such as body skeletons from inputs like images and videos. Over the past decade, this field has attracted growing interest and has been applied across various domains (Zheng, 2023). Deep learning-based method has significantly outperformed traditional methods in Human Pose Estimation, primarily owing to three key factors: the abundance of large-scale datasets, the enhanced representation power of deep neural networks, and the advent of high-performance hardware such as Graphics Processing Unit (GPU) platforms (Lan, 2022). It stands out in its ability to process sports analysis, computer-generated image, and so on, thereby simplifying tasks like action recognition and human tracking which emphasizes its potential as a pivotal tool in computer vision.

2D HPE refers to the information of human pose on two-dimensional image plane inferred from image or video. The main tasks include detecting key points of the human body (such as the head, shoulders, elbows, wrists, knees, etc.) and estimating their position and connection in the image. This approach typically does not take depth information into account and instead focuses on locating the location of key points in the human body in a two-dimensional image (Ben, 2021).

Beyond 2D human pose estimation research, experimenters' work extends to 3D HPE technology. 3D HPE refers to the estimation of human pose information in 3D space from images or videos. Different from two-dimensional pose estimation, three-dimensional pose estimation not only focuses on the position of key points in the image plane, but also considers the three-dimensional coordinates of these key points in space. Its main goal is to restore the posture of the human body in three-dimensional

^a <https://orcid.org/0009-0008-6166-6726>

^b <https://orcid.org/0009-0002-6220-4079>

^c <https://orcid.org/0009-0004-0731-1001>

space, including joint Angle, body orientation, spatial position of the body and other information. Therefore, 3D pose estimation may be more challenging than 2D (Ben, 2021).

To date, a wide range of research has been conducted by scientists to explore the vast potential of 2D HPE technology in computer vision. In 2021, experimenters discuss the motion recognition algorithm based on 2D method as well as using convolutional neural network (CNN) for action classification training (Chen, 2021). The results show that the motion recognition algorithm based on 2D HPE can improve the accuracy of motion recognition to a certain extent and show its potential and advantages in practical applications. Apart from 2D HPE research, experimenters have expanded their efforts into 3D HPE technology. In 2021, an experimental team use a red-green-blue Depth Image (RGBD) camera for 3D HPE and proved that this method can help the system reconstruct the human body's posture in 3D space more accurately (Fang, 2021). A series of related approaches have been tested and validated on publicly available data sets and demonstrated superior pose estimation results for 3D HPE.

In this paper, the goal is to present a comprehensive review for both 2D and 3D HPE based on deep learning and research papers from 2015 to 2024 are covered. The 2D human pose estimation calibration and key point visibility classification technique, as well as a Wi-Fi-based technique are introduced. In further discussion, this paper deeply explores the 3D HPE technology based on self-supervised learning and time encoder and regression decoder. For the experimental section, the data sets used to evaluate the various methods are detailed. A variety of evaluation metrics are proposed to evaluate the performance of different methods and attain straightforward conclusion. Finally, through the comparison of data analysis and experimental results, this paper discusses the advantages and limitations of the various methods and explore possible challenges and solutions in future research.

2 METHOD

2.1 WiFi Signals to Estimate 2D Human Poses

2.1.1 Method Introduction

In this experiment, 2D estimation of human body is performed through Wireless Fidelity (Wi-Fi), but the

information contained in Wi-Fi signals is limited, so it is necessary to extend the channel state information (CSI) sequence and integrate spatial information and dynamic information by using an air-frequency encoder. To enable the model to focus on specific channels, an evolutionary attention module is designed. The model outperforms the current state-of-the-art methods in PCK@20 by at least 16%.

2.1.2 Method Analysis

Wi-Fi signals to estimate 2D human poses can capture human posture without wearing devices, which can be widely used and promoted because it is based on Wi-Fi. But precision is harder to guarantee than sensors and cameras. In the processing of the human posture room, a large amount of data needs to be processed, so it will increase the cost and complexity.

2.2 3D Human Pose Estimation Based on Self-Supervision

2.2.1 Method Introduction

The method adopts Pose ResNet convolutional neural architecture, which is based on ResNet architecture. It is divided into 2D Pose CNN and 3D Pose CNN. The two modules are composed of feature extraction module and critical joint detection module. The model can achieve self-supervised learning and construct supervised information. The mechanism adopted is Convolutional Block Attention Module (CBAM), which can select the more important information from all the information and suppress the useless information. Using the CBAM module, the information of the image can be obtained, and the field of perception can be expanded, thus making the detection more accurate.

2.2.2 Method Analysis

The advantage of this method is that 3D data is not required, and 2D data can be used for training. Strong generalization ability. Reliance on external monitoring signals is reduced through self-monitoring. This method can also be used in combination with other technologies and has high scalability. However, the performance of this method in high-precision 3D attitude estimation is weaker than that of traditional supervised learning methods. If the data in the training data is biased, the self-monitoring method will also be affected.

2.3 Three-Dimensional Human Pose Estimation

2.3.1 Method introduction

The method uses Skinned Multi-Person Linear (SMPL) to evaluate and model human posture. The model is parameterized based on linear algebra. It uses low-dimensional parameters to display information about human posture and shape. The framework of this method is divided into three parts, which are linear encoding of data, time encoder and motion regression decoder. Inertial Measurement Unit (IMU) Feature Extractor enhances contract and time capture by combining two consecutive frames. Pose Triplet: Pose Triplet to transform the 2d pose sequence into 3d output. The advantage is that it does not rely on 3d data. The disadvantage is that 3d poses cannot be directly estimated from 2d images. Approach based on the encoder-decoder framework: Individual frames are transformed into feature vectors by Temporal Convolutional Encoder, Kinematic Regression Decoder is used to predict the motion of an object.

2.3.2 Method Analysis

Compared with the two-dimensional method, Three-Dimensional Human Pose Estimation is more abundant and accurate, which can make the generated figures more real and has been widely used now. However, the complexity of this method is high, the demand for data is large, and the propagation error will occur because of the reconstruction of three-dimensional information.

3 EXPERIMENTS

This section describes the more widely used 2D and 3D datasets by 2024, as well as the common evaluation metrics used for the different methods. The performance results of these methods on different datasets will also be presented.

3.1 Datasets

3.1.1 2D HPE Datasets

Deep learning-based methods are typically applied to large amounts of training data, and this part will introduce four mainly used 2D human pose datasets, which are Max Planck Institute for Informatics (MPII) (Andriluka,2014), Microsoft Common Objects in Context (COCO) (Lin,2014), PoseTrack (Andriluka,2018), Human-Art (Ju, 2023).

MPII dataset is a dataset to provide a high-quality and diverse source of human pose annotations including neck, head, elbow, which is important in HPE.

COCO Dataset is the most widely used 2D dataset including 200,000 labelled subjects with 80 categories such as people, animals and 330,000 images.

PoseTrack dataset is the datasets applied in the video HPE task, with massive keypoints tracking. Different scenarios include in the dataset such as human stand overlapping resulting in occlusion of body parts. Each human body is labeled with 15 major keypoints that describe the structure of human body.

Human-Art is also widely used in 2D methods. Comparison between these datasets is presented in Table 1.

Table 1: 2D HPE Datasets. S: Single-person M: Multi-person.

Image-based datasets						
Name	Year	S/M	Joints	Number of images		Evaluation protocol
				Train	Test	
MPII (Andriluka,2014)	2014	M	16	29k	12 k	PCKh mAp
		S	16	3.8k	1.7 k	
COCO2016 (Lin,2014)	2016	M	17	45k	80 k	AP
COCO2017 (Lin,2014)	2017	M	17	64 k	40 k	AP
Human-Art (Ju, 2023)	2023	M	17	10k	2k	mAP
Video-based datasets						
PoseTrack2017 (Andriluka,2018)	2017	M	15	250	214	mAP
PoseTrack2018 (Andriluka,2018)	2018	M	15	593	375	mAP

Table 2: 3D HPE Datasets. S: Single-person M: Multi-person.

Dataset	Year	Capture system	Environment	Including	S	M
AGORA (Patel,2021)	2021	High-quality textured scans	Indoor	14K images, 173K individual	Yes	Yes
DensePose-COCO (Güler, 2018)	2018	Manual annotation alignment with 3D models	Indoor	50K images, 5M correspondences	Yes	Yes
MuPoTS-3D (Mehta,2018)	2018	Multi-View Marker-less	Indoor and outdoor	8 subjects, 8k frames	Yes	Yes
Human3.6M (Ionescu,2014)	2014	Vicon	Indoor	11 subjects 3.6M frames	Yes	No

3.1.2 3D HPE Datasets

In 3D HPE datasets, Vicon and MoCap are required to be able to accurately acquire 3D coordinates.

Human3.6 M (Ionescu, 2014) is the dataset covers 17 daily activities observed from 4 different viewpoints. The dataset also utilizes a Vicon optical motion capture system to acquire high-precision 3.6 million 3D pose data. It is the most important for 3D HPE.

MuPoTS-3D (Multi-person Pose Tracking in 3D) dataset is a 3D dataset specialized for multi-person pose estimation and tracking. It contains 5 indoor scenes and 15 outdoor scenes.

DensePose-COCO (Güler, 2018) and AGORA (Patel, 2021) also included. Comparison between these datasets is presented in Table 2.

3.2 Evaluation Metrics

3.2.1 2D HPE Evaluation Metrics

Object Keypoint Similarity (OKS) is a mertric commonly used in COCO datasets to measures the similarity between predicted keypoints and ground truth keypoints.

The Average Precision (AP) metric is a measure of the accuracy of critical point detection based on precision and recall.

Percentage of Correct Keypoints (PCK) is used to measures 2D HPE method by checking if predicted keypoints fall within a threshold distance from ground truth keypoints, commonly 50% of the head segment length (PCKh@0.5).

3.2.2 3D HPE Evaluation Metrics

Area Under the Curve (AUC) plots the curve and calculates the area under the curve by calculating the keypoint correctness at multiple 3D PCK thresholds. It provides a summary of the overall performance of the model under different thresholds.

MPJPE (Mean Per Joint Position Error) is using the Euclidean distance to estimate the accuracy of the method as follows: J_i is the truth keypoints, J_i^* is the estimated keypoints.

$$MBJPE = \frac{1}{N} \sum_{i=1}^N \|J_i - J_i^*\|_2 \quad (1)$$

PA-MPJPE, is the MPJPE, where Procrustes alignment is used to remove global rotation, translation between the predicted and ground truth 3D position to reduce the impact form the global factors.

3.3 Performance for Each Method

3.3.1 2D Method

Table 3 shows the results and performance of many 2D methods on the COCO dataset, analyzed by comparing their AP values, backbone settings, and input sizes. The different 2D HPE methods are categorized into two types: top-down and bottom-up to compare the performance of the two methods. In general, the top-down method has better performance than the bottom-up method in terms of accuracy of pose estimation because the top-down method first identifies all the people in the scene and then predicts the location of the keypoints for each of them using the single-person HPE method. Since top-down methods in this category do not suffer from performance degradation due to high background complexity and high similarity between background and people. Moreover, bottom-up methods perform better and spend less time to finish the HPE task, compare with top-down methods, because bottom-up methods are advanced keypoint detection for the characters in the scene and then use keypoints association strategies such as Part Affinity Fields, Pose Residual Network to categorize them into single poses. The use of a transformer backbone in the top-down approach improves the efficiency of the HPE task without drastically decreasing its accuracy. In the bottom-up method, the use of the HRNet-W32 backbone provides the best performance

Table 3: 2D Methods Performance in COCO Datasets with AP Measure. T: Top-down; B: Bottom-up.

Performance Data on COCO									
	Year	Method	Backbone	Input size	AP.5	AP(M)	AP(L)	Params(M)	FLOPs(G)
T	2021	(Yang,2021)	HRNet-W48	256x192	92.2	71.3	81.1	17.5	21.8
	2022	(Ma,2022)	Transformer	256x192	89.8	71.7	82.1	20.8	8.7
B	2020	(Jin,2020)	Hourglass	512x512	85.1	62.7	74.6	-	-
	2021	(Luo,2021)	HRNet-W48	640x640	90.7	67.8	77.7	63.8	154.6
	2022	(Wang,2022)	HRNet-W32	640x640	91.2	68.3	79.3	-	-

Table 4: 3D Methods Performance on the Human3.6M Dataset.

Skeleton-only methods					MPJPE	PA-MPJPE
Approaches	Year	Method	Input	Params(M)	Average_1	Average_2
Direct	2017	(Pavlakos,2017)	Image	-	71.9	52.0
	2018	(Pavlakos,2018)	Image	-	56.3	41.7
Lifting	2020	(Wang,2020)	Video	-	42.5	32.7
	2022	(Zhang,2022)	Video	41	40.9	32.5
Human mesh recovery methods					MPJPE	PA-MPJPE
Mesh Output	2021	(Lin,2021)	Image	230	51.2	34.5
	2022	(Cho,2022)	Image	49	52.1	33.6
	2020	(Kocabas,2020)	Video	59	65.5	41.2
	2021	(Choi,2021)	Video	123	62.3	41.1

improvement, but is still lower than the top-down method.

3.3.2 3D Method

Table 4 shows the results and performance of 3D single-person HPE methods applied to the Human3.6M dataset and categorized into two groups. Almost all Single-person HPE methods are able to accurately estimate the 3D human pose on Human3.6M. However, they are affected by the limitations of the Human3.6M dataset, such as the lack of rich number and variety of character movements, actor body types and environment samples. When these methods are applied to estimate 3D human poses in more complex field scenarios, performance degradation problems occur. Meanwhile, estimating 3D human pose in video is superior to estimating human pose estimation through images, which exists with smaller MPJPE and higher accuracy.

For Skeleton-only Methods, these methods perform well with lower complexity, while the video input 2D-to-3D lifting methods outperform the direct estimation method overall by taking the support of a high-performance 2D pose detector. For Human mesh recovery methods, these methods suffer from high Mean Per Joint Position Error (MPJPE) by using the SMPL to recover, probably because of their high mesh recovery complexity, which requires simultaneous optimization and regression of multiple

objectives such as pose, shape, and surface details. This complexity may lead to the gradual accumulation of errors during the mesh reconstruction process, which results in high MPJPE. MPJPE is a limitation in evaluating such methods.

4 DISCUSSION

In 2D multi-person HPE, top-down methods have higher accuracy but are slower, while bottom-up methods are faster but relatively less accurate. To balance speed and accuracy, a hierarchical detection strategy can be adopted. First, the bottom-up method is used to quickly detect the approximate positions of all people, and then the top-down method is used to finely locate the key points for each person. Additionally, parallel computing techniques can be used in both bottom-up and top-down methods, accelerating processing speed through multi-GPU or distributed computing. Finally, the key point association strategy in bottom-up detection can be improved, for example, by using graph neural networks (GNNs) for key point association, enhancing the accuracy of key point assembly and thus improving overall accuracy.

The MPJPE of 3D skeleton and mesh recovery methods is relatively high, primarily because these methods regress both pose parameters and shape parameters simultaneously, leading to error

accumulation. To reduce error accumulation, a multi-task learning framework can be adopted to separately handle pose parameters and shape parameters. By sharing some network layers, the model can complement each other during the learning process, improving overall accuracy. For example, a joint network for pose estimation and shape reconstruction can be designed, with each having its own independent loss function. Additionally, the mesh reconstruction algorithm can be improved to reduce regression errors. Finally, a Generative Adversarial Network (GAN) model can be used to improve the quality of reconstructed meshes.

5 CONCLUSIONS

This paper introduces the current mainstream methods from 2D and 3D respectively. The two-dimensional human body is estimated based on Wi-Fi signal, and the CSI sequence is integrated by spatial encoder. Based on self-supervised 3D human Pose estimation, the Pose ResNet convolutional neural architecture is divided into 2D and 3D modules to detect features and key joints. 3D human pose estimation based on encoder and regression decoder enhances the accuracy of human pose estimation by adding both time and space.

At present, 2D motion capture can be divided into camera-based visual capture, hand-drawn animation, motion capture software, sensor capture. These methods have their own characteristics and use cases. In future research, it can be optimized for real-time interaction. Optimize the real-time performance of 2D motion capture, improve the interactivity of the system, and enable users to operate and edit more conveniently.

At present, the mainstream 3D motion capture method can be divided into optical camera systems, inertial measurement unit, depth cameras, optical tracking system, laser scanning system. These methods have been widely used in different fields. In future research, it can be combined with visual data, inertial data, depth data, etc., so as to enhance the accuracy of capture.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

- Ben Gamra, M., & Akhloufi, M. A., 2021. A review of deep learning techniques for 2D and 3D human pose estimation. *Image and Vision Computing*, 114, 104282. <https://doi.org/10.1016/j.imavis.2021.104282>
- Fang, Z., Wang, A., Bu, C., & Liu, C., 2021. 3D Human Pose Estimation Using RGBD Camera. 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI). <https://doi.org/10.1109/cei52496.2021.9574486>
- Lan, G., Wu, Y., Hu, F., & Hao, Q., 2022. Vision-Based Human Pose Estimation via Deep Learning: A Survey. *IEEE Transactions on Human-Machine Systems*, 1–16. <https://doi.org/10.1109/thms.2022.3219242>
- Yu, C., Chen, W., Li, Y., & Chen, C., 2021. Action Recognition Algorithm based on 2D Human Pose Estimation Method. *IEEE Xplore*. <https://doi.org/10.23919/CCC52363.2021.9550204>
- Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Nasser Khehtarnavaz, & Shah, M., 2023. Deep Learning-Based Human Pose Estimation: A Survey. <https://doi.org/10.1145/3603618>
- Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B., 2014. 2D human pose estimation: New benchmark and state of the art analysis. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 3686–3693. <https://doi.org/10.1109/CVPR.2014.471>
- Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., & Schiele, B., 2018. *PoseTrack: A Benchmark for Human Pose Estimation and Tracking*. *IEEE Xplore*. <https://doi.org/10.1109/CVPR.2018.00542>
- Ju, X., Zeng, A., Wang, J., Xu, Q., & Zhang, L., 2023. Human-Art: A versatile human-centric dataset bridging natural and artificial scenes. *arXiv preprint*. <https://doi.org/10.48550/arxiv.2303.02760>
- Cho, J., Youwang, K., & Oh, T.-H., 2022. *Cross-Attention of Disentangled Modalities for 3D Human Mesh Recovery with Transformers*. *ArXiv.org*. <https://arxiv.org/abs/2207.13820>
- Choi, H., Moon, G., Chang, J. Y., & Lee, K. M., 2020. *Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video*. *ArXiv.org*. <https://arxiv.org/abs/2011.08627>
- Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C., 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 1325–1339. <https://doi.org/10.1109/TPAMI.2013.248>
- Jin, S., Liu, W., Xie, E., Wang, W., Qian, C., Ouyang, W., & Luo, P., 2020. *Differentiable Hierarchical Graph Grouping for Multi-Person Pose Estimation*. *ArXiv.org*. <https://arxiv.org/abs/2007.11864>
- Kocabas, M., Athanasiou, N., & Black, M. J., 2020. *VIBE: Video Inference for Human Body Pose and Shape Estimation*. *ArXiv.org*. <https://doi.org/10.48550/arXiv.1912.05656>

- Lin, K., Wang, L., & Liu, Z., 2021. Mesh Graphormer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. <https://doi.org/10.1109/iccv48922.2021.01270>
- Luo, Z., Wang, Z., Huang, Y., Tan, T., & Zhou, E., 2020. Rethinking the Heatmap Regression for Bottom-up Human Pose Estimation. *ArXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2012.15175>
- Ma, H., Wang, Z., Chen, Y., Kong, D., Chen, L., Liu, X., Yan, X., Tang, H., & Xie, X., 2022. PPT: token-Pruned Pose Transformer for monocular and multi-view human pose estimation. *ArXiv.org*. <https://arxiv.org/abs/2209.08194>
- Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., & Black, M. J., 2019. AMASS: Archive of Motion Capture as Surface Shapes. <https://doi.org/10.1109/iccv.2019.00554>
- Mehta, D., Oleksandr Sotnychenko, Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., & Theobalt, C., 2018. Single-Shot Multi-person 3D Pose Estimation from Monocular RGB. <https://doi.org/10.1109/3dv.2018.00024>
- Patel, P., Huang, C.-H. P., Tesch, J., Hoffmann, D. T., Tripathi, S., & Black, M. J., 2021. AGORA: Avatars in Geography Optimized for Regression Analysis. *ArXiv.org*. <https://arxiv.org/abs/2104.14643>
- Pavlakos, G., Zhou, X., & Daniilidis, K., 2018. Ordinal Depth Supervision for 3D Human Pose Estimation. *ArXiv.org*. <https://arxiv.org/abs/1805.04095>
- Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K., 2017. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://doi.org/10.1109/cvpr.2017.139>
- Tang, W., Yu, P., & Wu, Y., 2018. Deeply Learned Compositional Models for Human Pose Estimation. *Lecture Notes in Computer Science*, 197–214. https://doi.org/10.1007/978-3-030-01219-9_12
- Wang, H., Zhou, L., Chen, Y., Tang, M., & Wang, J., 2022. Regularizing Vector Embedding in Bottom-Up Human Pose Estimation. *Lecture Notes in Computer Science*, 107–122. https://doi.org/10.1007/978-3-031-20068-7_7
- Wang, J., Yan, S., Xiong, Y., & Lin, D., 2020. Motion Guided 3D Pose Estimation from Videos. *ArXiv.org*. <https://arxiv.org/abs/2004.13985>
- Yang, S., Quan, Z., Nie, M., & Yang, W., 2021. TransPose: Keypoint Localization via Transformer. *ArXiv.org*. <https://doi.org/10.48550/arXiv.2012.14214>
- Zhang, J., Tu, Z., Yang, J., Chen, Y., & Yuan, J., 2022. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://doi.org/10.1109/cvpr52688.2022.01288>
- Güler, R. A., Neverova, N., & Kokkinos, I., 2018. DensePose: Dense Human Pose Estimation In The Wild. *ArXiv:1802.00434 [Cs]*. <https://arxiv.org/abs/1802.00434>
- Kocabas, M., Athanasiou, N., & Black, M. J., 2020. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Choi, H., Moon, G., Chang, J. Y., & Lee, K. M., 2021. Beyond static features for temporally consistent 3D human pose and shape from a video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Bao, W., Ma, Z., Liang, D., Yang, X., & Niu, T., 2023. Pose ResNet: 3D Human Pose Estimation Based on Self-Supervision. *Sensors (Basel, Switzerland)*, 23(6), 3057. <https://doi.org/10.3390/s23063057>
- Liao, X., Dong, J., Song, K., & Xiao, J., 2023. Three-Dimensional Human Pose Estimation from Sparse IMUs through Temporal Encoder and Regression Decoder. *Sensors (Basel, Switzerland)*, 23(7), 3547. <https://doi.org/10.3390/s23073547>